

Proofs

Proposition 1. Let (f, Σ) be a constrained decision-function and $\mathbf{x} \in [\Sigma]$ be an instance s.t. $f(\mathbf{x}) = 1$ (resp. $f(\mathbf{x}) = 0$).

- The (weak) contrastive explanations for \mathbf{x} given (f, Σ) are the sets of literals occurring in the implicates of $\forall \mathbf{x} \cdot (\Sigma \Rightarrow f)$ (resp. $\forall \mathbf{x} \cdot (\Sigma \Rightarrow \bar{f})$).
- The (subset-minimal) contrastive explanations for \mathbf{x} given (f, Σ) are the sets of literals occurring in the prime implicates of $\forall \mathbf{x} \cdot (\Sigma \Rightarrow f)$ (resp. $\forall \mathbf{x} \cdot (\Sigma \Rightarrow \bar{f})$).
- The (minimum-size) contrastive explanations for \mathbf{x} given (f, Σ) are the sets of literals occurring in the minimum-size prime implicates of $\forall \mathbf{x} \cdot (\Sigma \Rightarrow f)$ (resp. $\forall \mathbf{x} \cdot (\Sigma \Rightarrow \bar{f})$).

Proof. The proof of Proposition 1 from [17] shows that the (weak, resp. subset-minimal) abductive explanations for \mathbf{x} given (f, Σ) are the sets of literals occurring in the implicants (resp. prime implicants) t of $\Sigma \Rightarrow f$ such that $t \subseteq t_{\mathbf{x}}$ when $f(\mathbf{x}) = 1$. As a direct consequence, we also have that the (weak, resp. subset-minimal) abductive explanations for \mathbf{x} given (f, Σ) are the sets of literals occurring in the implicants (resp. prime implicants) of $\Sigma \Rightarrow \bar{f}$ such that $t \subseteq t_{\mathbf{x}}$ when $f(\mathbf{x}) = 0$. Then, we take advantage of the notion of universal literal quantification considered in [14] and use Proposition 11 from [13] to get that the sets of literals occurring in the implicates (resp. prime implicates) of $\forall \mathbf{x} \cdot (\Sigma \Rightarrow f)$ are the (weak, resp. subset-minimal) contrastive explanations for \mathbf{x} given (f, Σ) when $f(\mathbf{x}) = 1$, and that the sets of literals occurring in the implicates (resp. prime implicates) of $\forall \mathbf{x} \cdot (\Sigma \Rightarrow \bar{f})$ are the (weak, resp. subset-minimal) contrastive explanations for \mathbf{x} given (f, Σ) when $f(\mathbf{x}) = 0$. Finally, the previous result about subset-minimal contrastive explanations extend to minimum-size contrastive explanations, given that the minimum-size contrastive explanations for \mathbf{x} given (f, Σ) are the subset-minimal contrastive explanations for \mathbf{x} given (f, Σ) that are of minimum size. \square

Proposition 2. Let $F \in \text{RF}_n$ be a random forest and $\mathbf{x} \in \{0, 1\}^n$ be an instance. The number of minimum-size contrastive explanations for \mathbf{x} given $(F, 1)$ can be exponential in the number n of attributes.

Proof. Let $k = \lfloor \frac{n}{2} \rfloor$. Consider the DNF formula $f = \bigvee_{i=0}^{k-1} (x_{2i+1} \wedge x_{2i+2})$ and the instance $\mathbf{x} \in \{0, 1\}^n$ such that $x_i = 1$ for each $i \in [n]$. We have $\forall \mathbf{x} \cdot f \equiv f$. The subset-minimal contrastive explanations for \mathbf{x} given $(f, 1)$ are the sets of literals occurring in the prime implicates of $\forall \mathbf{x} \cdot f$, thus the sets of literals occurring in the prime implicates of f . They all have the same size (k), hence they are all minimal-size contrastive explanations for \mathbf{x} given $(f, 1)$. Consider now a random forest F from RF_n equivalent to f (see Proposition 2 from [4] for the generation of F in polynomial time from f). The fact that f has 2^k prime implicates completes the proof. \square

Proposition 3. Let (f, Σ) be a constrained decision-function and $\mathbf{x} \in [\Sigma]$ be an instance such that $f(\mathbf{x}) = 1$ (resp. $f(\mathbf{x}) = 0$). \mathbf{x} has a (weak) contrastive explanation given (f, Σ) if and only if $\neg f \wedge \Sigma$ (resp. $f \wedge \Sigma$) is satisfiable.

Proof. Suppose that $\mathbf{x} \in [\Sigma]$ is such that $f(\mathbf{x}) = 1$. If $\bar{f} \wedge \Sigma$ is unsatisfiable then there is no model \mathbf{x}' of Σ that is a model of \bar{f} , hence \mathbf{x} does not have any (weak) contrastive explanation given (f, Σ) . Similarly, if $\mathbf{x} \in [\Sigma]$ is such that $f(\mathbf{x}) = 0$ and $f \wedge \Sigma$ is unsatisfiable then there is no model \mathbf{x}' of Σ that is a model of f , hence \mathbf{x} does not have any (weak) contrastive explanation given (f, Σ) . \square

Proposition 4. Let (f, Σ) be a constrained decision-function and $\mathbf{x} \in [\Sigma]$ be an instance. Deciding whether \mathbf{x} has a (weak) contrastive explanation given (f, Σ) is NP-complete. NP-hardness still holds when f is represented by a random forest from RF_n and $\Sigma = 1$.

Proof.

- **Membership to NP:** The following nondeterministic algorithm runs in time polynomial in the input size (\mathbf{x} and a representation of f): Guess $c \subseteq t_{\mathbf{x}}$ and check (in deterministic polynomial time) that $\mathbf{x}_c \models \Sigma$ and $f(\mathbf{x}_c) \neq f(\mathbf{x})$.
- **NP-hardness:** we prove that the restriction of the decision problem when f is a random forest F from RF_n and $\Sigma = 1$ is NP-hard by reduction from the satisfiability problem for CNF formulae. Let $\alpha = c_1 \wedge \dots \wedge c_m$ be a CNF formula over X_n . Let \mathbf{x} be any instance from $\{0, 1\}^n$ such that $\{\bar{\ell} : \ell \in c_1\} \subseteq t_{\mathbf{x}}$. We associate with α in polynomial time the pair $\langle \mathbf{x}, F \rangle$, where F is a random forest from RF_n equivalent to $\neg \alpha$ (see Proposition 2 from [4] for the generation of F). We have $F(\mathbf{x}) = 1$ since $\alpha(\mathbf{x}) = 0$ by construction of \mathbf{x} . Now, from Proposition 3, deciding whether \mathbf{x} has a (weak) contrastive explanation given $(F, 1)$ amounts to deciding whether $\neg F$ is satisfiable, thus to deciding whether α is satisfiable. This concludes the proof. \square

Proposition 5. Let (f, Σ) be a constrained decision-function and $\mathbf{x} \in [\Sigma]$ be an instance. Let $c \subseteq t_{\mathbf{x}}$. Deciding whether c is a (weak) contrastive explanation for \mathbf{x} given (f, Σ) is in P.

Proof. By definition, c is a (weak) contrastive explanation for \mathbf{x} given (f, Σ) if and only if $\mathbf{x}_c \in [\Sigma]$ and $f(\mathbf{x}_c) \neq f(\mathbf{x})$. Both tests can be achieved in polynomial time since \mathbf{x}_c is an interpretation over X_n and f and Σ are built upon X_n . \square

Proposition 6. Let (f, Σ) be a constrained decision-function and $\mathbf{x} \in [\Sigma]$ be an instance. Let $c \subseteq t_{\mathbf{x}}$. Deciding whether c is a (subset-minimal) contrastive explanation for \mathbf{x} given (f, Σ) is coNP-complete. coNP-hardness still holds when f is represented by a random forest from RF_n and $\Sigma = 1$.

Proof.

- **Membership to coNP:** we first check that c is a (weak) contrastive explanation for \mathbf{x} given (f, Σ) . This is done in polynomial time (see Proposition 5). Now, c is not a (subset-minimal) contrastive explanation for \mathbf{x} given (f, Σ) if and only if there exists a proper subset c' of c that is a (weak) contrastive explanation for \mathbf{x} given (f, Σ) . Deciding whether such a c' exists is in NP: it is enough to guess $c' \subset c$ and to test in deterministic polynomial time that c' is a (weak) contrastive explanation for \mathbf{x} given (f, Σ) .
- **coNP-hardness:** we prove that the restriction of the decision problem when f is a random forest F from RF_n and $\Sigma = 1$ is coNP-hard by reduction from the minimal model checking problem for CNF formulae, which is coNP-complete [9]. The latter problem is as follows:

- **Input:** an instance $\mathbf{x} \in \{0, 1\}^n$ and a CNF formula $\alpha = \bigwedge_{i=1}^m c_i$ over X_n .
- **Question:** Is $t_{\mathbf{x}}$ a subset-minimal model of α , i.e., a model of α such that the set of positive literals in $t_{\mathbf{x}}$ is minimal w.r.t. set-inclusion?

We first prove the following lemma:

Lemma 1. *Let $\mathbf{x} \in \{0, 1\}^n$ and $\alpha = \bigwedge_{i=1}^m c_i$ a CNF formula over X_n . Let \mathbf{z} be the instance of $\{0, 1\}^{n+1}$ such that $z_i = 0$ for $i \in [n+1]$. Let $c = \{\bar{x}_i : x_i = 1, i \in [n]\} \cup \{\bar{x}_{n+1}\}$. c is a (weak) contrastive explanation for \mathbf{z} given $(\bar{\alpha} \vee \bar{x}_{n+1}, 1)$ if and only if $t_{\mathbf{x}}$ is a model of α . Furthermore, \mathbf{z}_c coincides with \mathbf{x} over X_n .*

Proof. We have that $t_{\mathbf{z}}$ is a model of $\bar{\alpha} \vee \bar{x}_{n+1}$ since $\bar{x}_{n+1} \in t_{\mathbf{z}}$. We also have $c \subseteq t_{\mathbf{z}}$ since c contains only negative literals. c is a (weak) contrastive explanation for \mathbf{z} given $(\bar{\alpha} \vee \bar{x}_{n+1}, 1)$ if and only if \mathbf{z}_c is a model of $\alpha \wedge x_{n+1}$. Since $c = \{\bar{x}_i : x_i = 1, i \in [n]\} \cup \{\bar{x}_{n+1}\}$, \mathbf{z}_c coincides with \mathbf{x} over X_n . Since, by construction, $x_{n+1} \in t_{\mathbf{z}_c}$, \mathbf{z}_c is a model of $\alpha \wedge x_{n+1}$ if and only if \mathbf{z}_c is a model of α if and only if $t_{\mathbf{x}}$ is a model of α . \square

Then the reduction from the minimal model checking problem is as follows: to any input $\langle \mathbf{x}, \alpha \rangle$ of the minimal model checking problem we associate in polynomial time the following triple $\langle F, \mathbf{z}, c \rangle$ where

$$F = \{T(\bar{c}_1 \vee \bar{x}_{n+1}), \dots, T(\bar{c}_m \vee \bar{x}_{n+1}), \underbrace{\top, \dots, \top}_{m-1}\}$$

is a random forest over X_{n+1} containing $2m - 1$ trees. Each $T(\bar{c}_i \vee \bar{x}_{n+1})$ ($i \in [m]$) is a decision tree over X_{n+1} equivalent to the formula $\bar{c}_i \vee \bar{x}_{n+1}$ (this tree can be generated in time linear in the size of c_i). By construction, F is equivalent to $\bar{\alpha} \vee \bar{x}_{n+1}$, so that \bar{F} is equivalent to $\alpha \wedge x_{n+1}$. Finally, take \mathbf{z} and c as given in Lemma 1. From Lemma 1, since \mathbf{z}_c coincides with \mathbf{x} over X_n and since every model of \bar{F} must set x_{n+1} to 1 as \mathbf{z}_c does it, $t_{\mathbf{x}}$ is a subset-minimal model of α if and only if c is a (subset-minimal) contrastive explanation for \mathbf{z} given $(F, 1)$. \square

Proposition 7. Let (f, Σ) be a constrained decision-function and $\mathbf{x} \in [\Sigma]$ be an instance. Let $c \subseteq t_{\mathbf{x}}$. Deciding whether c is a minimum-size contrastive explanation for \mathbf{x} given (f, Σ) is **coNP**-complete. **coNP**-hardness still holds when f is represented by a random forest from RF_n and $\Sigma = 1$.

Proof.

- **Membership to coNP:** we first check that c is a (weak) contrastive explanation for \mathbf{x} given (f, Σ) . This is done in polynomial time (see Proposition 5). Now, c is not a (minimum-size) contrastive explanation for \mathbf{x} given (f, Σ) if and only if there exists a subset c' of $t_{\mathbf{x}}$ that is a (weak) contrastive explanation for \mathbf{x} given (f, Σ) and is such that $|c'| < |c|$. Deciding whether such a c' exists is in **NP**: it is enough to guess $c' \subset t_{\mathbf{x}}$ and to test in deterministic polynomial time that c' is a (weak) contrastive explanation for \mathbf{x} given (f, Σ) and that $|c'| < |c|$.
- **coNP-hardness:** we prove that the restriction of the decision problem when f is a random forest F from RF_n and $\Sigma = 1$ is **coNP**-hard by reduction from the minimum-cardinality model checking for CNF formulae, which is as follows:
 - **Input:** an instance $\mathbf{x} \in \{0, 1\}^n$ and a CNF formula $\alpha = \bigwedge_{i=1}^m c_i$ over X_n .
 - **Question:** Is $t_{\mathbf{x}}$ a minimum-cardinality model of α , i.e., a model of α such that the set of positive literals in $t_{\mathbf{x}}$ is minimal w.r.t. cardinality?

We first prove that the minimum-cardinality model checking problem for CNF formulae is **coNP**-hard. The reduction is from **UNSAT**: starting with a CNF formula $\beta = \bigwedge_{i=1}^m c_i$ over X_n , let us associate in polynomial time the CNF formula $\alpha = \bigwedge_{j=i+1}^{2n+1} \bigwedge_{i=1}^m (x_j \vee c_i)$ over X_{2n+1} and the instance $\mathbf{x} \in \{0, 1\}^{2n+1}$ that sets every x_i ($i \in [n]$) to 0 and every x_i ($i \in [n+1, 2n+1]$) to 1. α contains $(n+1) \cdot m$ clauses and is of size $\mathcal{O}(|\beta|^2)$. If β is unsatisfiable, then α is equivalent to $\bigwedge_{i=n+1}^{2n+1} x_i$. Hence $t_{\mathbf{x}}$ is the sole minimum-cardinality model of α (it contains $n+1$ variables set to 1). If β is satisfiable, then it has a model over X_n , thus it also has a minimum-cardinality model over X_n , and this model contains at most n variables set to 1. Then the extension over X_{2n+1} of this minimum-cardinality model obtained by setting every x_i ($i \in [n+1, 2n+1]$) to 0 is a minimum-cardinality model of β . Since this model contains at most n variables set to 1, $t_{\mathbf{x}}$ is not a minimum-cardinality model of α . This concludes the **coNP**-hardness proof for the problem of checking a minimum-cardinality model for CNF formulae.

Then we reduce the minimum-cardinality model checking problem to the problem of deciding whether c is a (minimum-size) contrastive explanation for \mathbf{x} given $(F, 1)$ where $F \in \text{RF}_n$ is a random forest. To any input $\langle \mathbf{x}, \Sigma \rangle$ of the minimum-cardinality model checking problem we associate in polynomial time the triple $\langle F, \mathbf{z}, c \rangle$ as given in the proof of Proposition 6. We use Lemma 1 to conclude that $t_{\mathbf{x}}$ a minimum-cardinality model of Σ if and only if c is a (minimum-size) contrastive explanation for \mathbf{x} given $(F, 1)$. \square

Proposition 8. Let (f, Σ) be a constrained decision-function and $\mathbf{x} \in [\Sigma]$ be an instance such that $f(\mathbf{x}) = 1$.⁴ Let $(C_{\text{soft}}, C_{\text{hard}})$ be an instance of the **PARTIAL MAXSAT** problem such that $C_{\text{soft}} = t_{\mathbf{x}}$ and $C_{\text{hard}} = \text{CNF}(\Sigma \wedge \bar{f})$ where $\text{CNF}(\Sigma \wedge \bar{f})$ is a CNF encoding of $\Sigma \wedge \bar{f}$. Let \mathbf{z}^* be an optimal solution of $(C_{\text{soft}}, C_{\text{hard}})$. Then, $c = t_{\mathbf{x}} \setminus t_{\mathbf{z}^*}$ is a minimum-size contrastive explanation for \mathbf{x} given (f, Σ) and we have $t_{\mathbf{x}_c} = t_{\mathbf{z}^*} \cap L_{X_n}$.

Proof. Let \mathbf{z}^* be any optimal solution of $(C_{\text{soft}}, C_{\text{hard}})$. On the one hand, \mathbf{z}^* is a model of C_{hard} . Since $\text{CNF}(\Sigma \wedge \bar{f})$ is query-equivalent to $\Sigma \wedge \bar{f}$, $t_{\mathbf{z}^*} \cap L_{X_n}$ is a model of $\Sigma \wedge \bar{f}$. Now, since \mathbf{z}^* is an optimal solution of $(C_{\text{soft}}, C_{\text{hard}})$, \mathbf{z}^* satisfies a maximal number of soft clauses from C_{soft} . Since those soft clauses are precisely the literals occurring in $t_{\mathbf{x}}$, the set of literals $c = t_{\mathbf{x}} \setminus t_{\mathbf{z}^*}$ is a subset of literals of $t_{\mathbf{x}}$ of minimum-size such that $t_{\mathbf{x}_c} = t_{\mathbf{z}^*} \cap L_{X_n}$ is a model of $\Sigma \wedge \bar{f}$. Stated otherwise, c is a minimum-size contrastive explanation for \mathbf{x} given (f, Σ) . \square

Proposition 9. Let $F \in \text{RF}_n$ such that $F(\mathbf{x}) = 1$, $\mathbf{x} \in \{0, 1\}^n$, and $t \subseteq t_{\mathbf{x}}$. Deciding whether t is a minimum-size abductive explanation for \mathbf{x} given F is Π_2^P -complete.

Proof.

- **Membership to Π_2^P :** we show that the problem belongs to Π_2^P in the general case when the classifier f is a Boolean function $f \in \mathcal{F}_n$. To get the result, we prove that the complementary problem belongs to Σ_2^P . Then, in order to decide whether t is *not* a (minimum-size) abductive explanation for \mathbf{x} given f , it is enough to first test whether t is a (weak) abductive explanation for \mathbf{x} given f using one call to an **NP** oracle; if t is not such an explanation, then it

⁴ If \mathbf{x} is such that $f(\mathbf{x}) = 0$, then consider \bar{f} instead of f .

is not a (minimum-size) abductive explanation for \mathbf{x} given f ; in the remaining case, it is enough to guess $t' \subseteq t_{\mathbf{x}}$, to check that $|t'| < |t|$, and finally to check (using one call to an NP oracle) that t' is a (weak) abductive explanation for \mathbf{x} given f .

- Π_2^P -hardness: let us associate in polynomial time with $\langle \mathbf{x} \in \{0, 1\}^n, F = \{T_1, \dots, T_m\} \in \mathbf{RF}_n, k \leq n \rangle$ such that $F(\mathbf{x}) = 1$ the triple $\langle \mathbf{x}', F', t \rangle$ where $\mathbf{x}' \in \{0, 1\}^{n+k+1}$ coincides with \mathbf{x} on its first n coordinates and is such that $x'_j = 1$ for each $j \in [n+1, n+k+1]$, $F' = \{T'_1, \dots, T'_m\} \in \mathbf{RF}_{n+k+1}$ is such that T'_i ($i \in [m]$) is a decision tree equivalent to $T_i \vee \bigwedge_{j=n+1}^{n+k+1} x_j$. Clearly, each decision tree T'_i ($i \in [m]$) can be generated in time $\mathcal{O}(k \cdot |T_i|)$ given that the term $\bigwedge_{j=n+1}^{n+k+1} x_j$ can be represented by a decision tree containing k decision nodes and that a decision tree representing the disjunction of two decision trees can be computed in time in the product of the sizes of the two trees (see [29]). Since $k \leq n$, $|F'|$ is polynomial in $|\mathbf{x}| \cdot |F|$. By construction, F' is equivalent to $F \vee \bigwedge_{j=n+1}^{n+k+1} x_j$ so that $F'(\mathbf{x}') = 1$. Finally, let $t = \bigwedge_{j=n+1}^{n+k+1} x_j$. By construction, t is an implicant of F' such that $t \subseteq t_{\mathbf{x}'}$. t contains $k+1$ characteristics. t is a prime implicant of F' unless F is valid (in which case F' is valid as well). More precisely, if F is valid, then \top is the unique prime implicant of F' , else the prime implicants of F' are the prime implicants of F , plus t . So if F has a (minimum-size) abductive explanation for \mathbf{x} given F containing at most k characteristics, then this explanation is also a (minimum-size) abductive explanation for \mathbf{x}' given F' , showing that t is *not* a (minimum-size) abductive explanation for \mathbf{x}' given F' . In the remaining case, every (minimum-size) abductive explanation for \mathbf{x} given F contains at least $k+1$ characteristics (hence F is not valid). This shows that t is a (minimum-size) abductive explanation for \mathbf{x}' given F' , and this completes the proof.

□